# Introduction to Data Analysis (Data Analysis I)

Peter Green

Emeritus Professor of Statistics
University of Bristol

# Outline

1 1. Paradigm

2 2. Simple linear regression

3 3. Linear models

4 4. Model formulation

5 5. Least squares estimation

6 6. Statistical performance

7 7. Normal theory assumptions

8 8. Model choice in linear models

9 9. Factorial experiments

10 10. Residual analysis

11 11. Variable Selection

12 12. Weighted Least Squares

13 13. Data collection

# How does this course fit in the programme?

- This unit builds on Introduction to Probability and Statistics (Dario Domingo, September/October 2025) – we use the distribution theory and basics of inference to discuss Linear Models, the main technical aspect to this course
- It is followed up by Data Analysis II (Thierry Chekouo, May/June 2026) - which leads on from Linear Models to Generalized Linear Models

# Topics in syllabus, not yet in slides 1

- Ability to manage various data sets and perform exploratory data analysis.
  - Different data collection methods and data types with real life health-related examples e.g routine malaria surveillance data vs cohort studies.
- Know different approaches to data collection and the use of different data types.
  - Sampling techniques: Probability and non-probability techniques and examples.
- Know how to pose relevant questions in data analysis, and formulate these questions statistically.
  - Sample size determination for epidemiological studies
  - Limitations and advantages of different data types and when to use what data (i.e what is the question being asked).
  - Accessing data, review of metadata, using data dictionaries, study types.
  - Checking for basic errors and distributions; documentation of errors or missing data; creating variables and transforming data.
  - Univariate summary statistics, tabulations and descriptive figures.
  - Preparing statistical analysis plans

# Topics in syllabus, not yet in slides 2

- Understand the basics of systematic literature review and meta-analysis
  - What is systematic literature review? What types of questions does it answer? How is it done?
  - What is meta-analysis? What types of questions does it answer? How is it done?
  - Be able to interpret and use results from systematic literature review and meta-analysis with application to mathematical epidemiology using literature based examples

# 1. Paradigm for probabilistic statistical inference

When does “data analysis” become “statistical inference”? ... when we go beyond statements about the the data we have to those about the data we might have had or even what caused these data?

This leads us to distinguish our sample from the population.

The sample might have been different while the population is what we are really interested in

**Example (a finite population)**

We are interested in the population of 18–25 year-old people in Cameroon, and we study this by drawing a sample of such people, and analysing that instead. What does that sample tell us about the population?

**Example (an infinite population)**

We are interested in the accuracy of a new instrument for measuring .... What does that sample tell us about the population?

# The main ingredients, when the population is infinite

When the population is infinite, we can think of it as a probability distribution!

- Regard the data $x$ as the observed values of a random vector $X$
- Model the distribution of $X$ with a probability (density) function $f(X;\theta)$ depending on a parameter $\theta$
- Use the observed $x$ in $f(x;\theta)$ to make statements about $\theta$, e.g.
    - estimate $\theta$
    - calculate a confidence interval for $\theta$
    - test hypotheses about the value of $\theta$

# The main ingredients, when the population is infinite

When the population is infinite, we can think of it as a probability distribution!

- Regard the data $x$ as the observed values of a random vector $X$
- Model the distribution of $X$ with a probability (density) function $f(X;\theta)$ depending on a parameter $\theta$
- Use the observed $x$ in $f(x;\theta)$ to make statements about $\theta$, e.g.
  - estimate $\theta$
  - calculate a confidence interval for $\theta$
  - test hypotheses about the value of $\theta$

# 2. Simple linear regression – Example: Oxygen Uptake

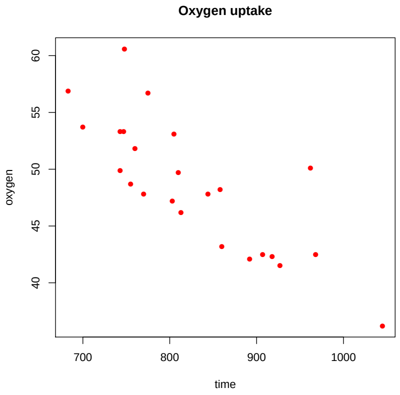
## Example

*For each of 24 males, the maximum volume of oxygen uptake in the blood and the time taken to run 2 miles (in seconds) were measured. Interest lies on the dependency between the time to run 2 miles and the oxygen uptake.*

```
oxy=c(42.3, 53.1 ,42.1, 50.1, 42.5, 42.5, 47.8, 49.9, 36.2, 49.7, 41.5, 46.2, 48.2, 43.2, 51.8, 53.3, 53.3, 47.2,


56.9, 47.8, 48.7, 53.7, 60.6, 56.7)
time=c(918, 805, 892, 962, 968, 907, 770, 743,1045, 810, 927, 813, 858, 860, 760, 747,743, 803, 683, 844,
755, 700, 748, 775)
plot(time, oxy, col="red", main="Oxygen uptake", type="p",pch=16)
```

# Simple linear regression: Oxygen Uptake Example



- For individual $i$, let $x_i$ be the time to run 2 miles, and $Y_i$ be the maximum volume of oxygen uptake, $i = 1, \cdots, 24$.
- A possible model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \; i = 1, \cdots, 24,$$

where $e_i$ are independent random variables with variance $\sigma^2$, and $\beta_0$ and $\beta_1$ are constants.

# Simple linear regression

We want to choose $\alpha$ and $\beta$ to minimise the sum of squares $SS = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$. Differentiating partially,

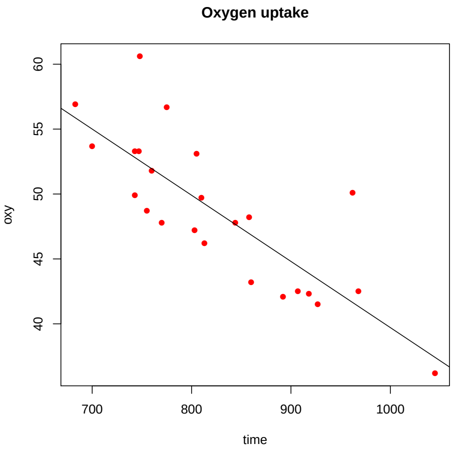
$$\frac{\partial SS}{\partial \alpha} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-1) = 0 \text{ if and only if } \sum_i y_i = n\alpha + \beta \sum_i x_i$$

$$\frac{\partial SS}{\partial \beta} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-x_i) = 0 \text{ if and only if } \sum_i x_i y_i = \alpha \sum_i x_i + \beta \sum_i x_i^2$$

Solving for $\alpha$ and $\beta$ we find the least squares estimates

$$\hat{\beta} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \text{ and } \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

# Simple linear regression: Oxygen Uptake Example



# Simple linear regression

Move this and next slide to be an example of the general case

Hence

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$\Rightarrow (\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Also

$$\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

# Simple linear regression

From which we can derive the following

$$\widehat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = s_x^2(n-1)$$

where $s_x^2$ is the sample variance of the observed $x_i, i = 1, \ldots, n$.

# Example

## Example

*Consider the relationship between the height $H$ and the weight $W$ of individuals in a certain city. Certainly there's no functional relationship between $H$ and $W$, but there does seem to be some kind of relation. We consider them as random variables and postulate that $(H, W)$ has a bivariate normal distribution. Then*

$$E[W|H = h] = \beta_0 + \beta_1 h$$

*where $\beta_0$ and $\beta_1$ are functions of the parameters in a bivariate normal density. Note that $\beta_0$, $\beta_1$ and $h$ are all constants. We may write*

$$W = \beta_0 + \beta_1 h + E$$

*where the error $E$ is a normally distributed random variable with mean zero.*

# Example (cont.)

## Example

*Thus if we observe the heights and weights of a sample of $n$ people, the model for the weights $w_i, \ldots, w_n$ is given by*

$$w_i = \beta_0 + \beta_1 h_i + e_i, \quad fori = 1, \ldots, n$$

*where the error $e_i \sim N(0, \sigma^2)$. This is a **simple linear regression model**: a linear model with one explanatory variable. In matrix form,*

$$\boldsymbol{w} = \boldsymbol{X}\beta + \boldsymbol{e}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 1 & h_1 \\ 1 & h_2 \\ \vdots & \vdots \\ 1 & h_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

# Simple linear regression: Example Oxygen uptake

move this and later oxygen example slides to correct place in the narrative

```
lm(formula = oxy ~ time)
```

```
                  Min         1Q      Median        3Q        Max
Residuals:
               -3.6461    -2.6422    -0.6792     1.0620     8.4545
Coefficients:
              Estimate    Std. Error    t value       Pr(>|t|)
 (Intercept)  90.69476     6.54324      13.861     2.38e-12 ***
 time         -0.05099     0.00787      -6.479     1.62e-06 ***
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.499 on 22 degrees of freedom
Multiple R-squared: 0.6561, Adjusted R-squared: 0.6405
F-statistic: 41.98 on 1 and 22 DF, p-value: 1.616e-06
```

# ANOVA: Example Oxygen uptake

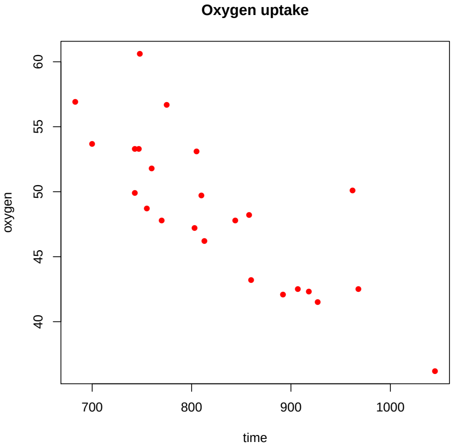
```
anova(lmoxy)
Analysis of Variance Table
Response: oxy
            Df   Sum Sq   Mean Sq   F value        Pr(>F)
  time       1   513.84    513.84    41.978   1.616e-06 ***
  Residuals 22   269.30     12.24
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Exercise: Oxygen Uptake

**Exercise**

*For the oxygen uptake data, test the significance of the slope parameter at the 0.1% significance level.*

# Simple linear regression: Example Oxygen uptake

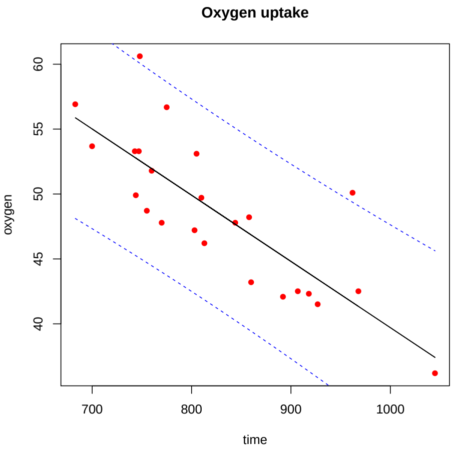


# Simple linear regression: Example Oxygen uptake

```
lm(formula = oxy ~ time)
```

```
              Min        1Q     Median        3Q       Max
Residuals:  -3.6461   -2.6422   -0.6792    1.0620    8.4545
Coefficients:
              Estimate    Std. Error    t value       Pr(>|t|)
 (Intercept)  90.69476      6.54324     13.861     2.38e-12 ***
_time         -0.05099      0.00787     -6.479     1.62e-06 ***
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.499 on 22 degrees of freedom
Multiple R-squared: 0.6561, Adjusted R-squared: 0.6405
F-statistic: 41.98 on 1 and 22 DF, p-value: 1.616e-06
```

# Prediction: Oxygen Uptake Exercise



# Prediction: Oxygen Uptake Exercise

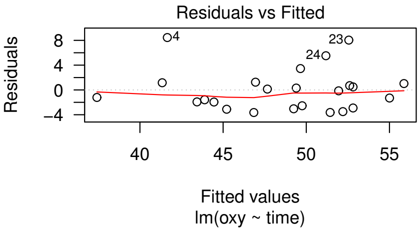
**Exercise**

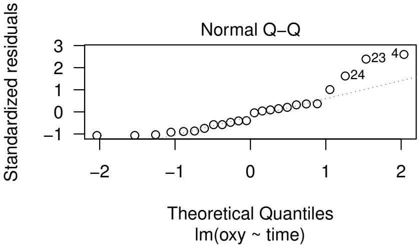
*For the oxygen uptake data, use the simple linear regression model to predict the maximum volume of oxygen uptake in the blood for for 3 individuals that take 750, 850 and 950 seconds to run 2 miles. Give a 95% confidence interval for these predictions.*

# ANOVA: Example Oxygen uptake

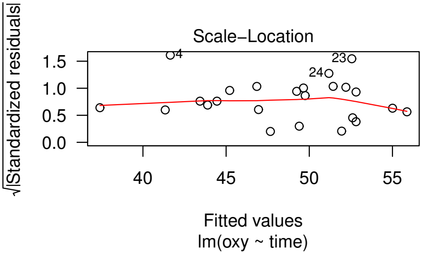
```
anova(lmoxy)
Analysis of Variance Table
Response: oxy
            Df   Sum Sq   Mean Sq   F value        Pr(>F)
  time       1   513.84    513.84    41.978   1.616e-06 ***
  Residuals 22   269.30     12.24
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

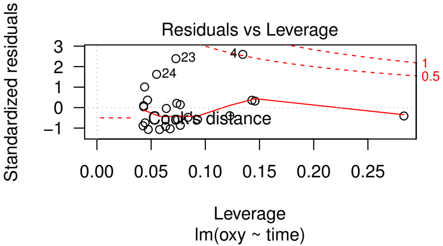
# Residuals Analysis: Oxygen uptake





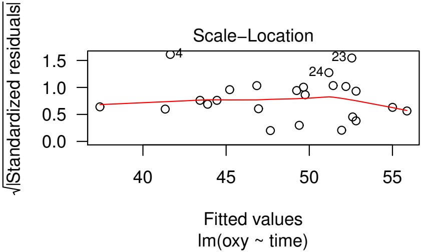
# Residuals Analysis: Oxygen uptake





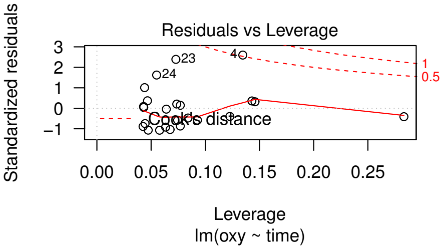
# Scale–location

Standardised residuals should be within (-2,2) range.



# Residuals versus Leverage

Tells us influential variables. If Cook's distance $D_i > 0.5$ we must be concerned and if $D_i > 1$ check the variable.



# 3. Linear models: Introduction

Linear models are statistical models where the expected response is a linear function of parameters – it is a very large class of models that includes many simpler models (such as linear regression and analysis of variance) as important special cases

- systematic treatment
  - model formulation in various ways
  - least squares estimation
  - optimality of least squares
  - connection with maximum likelihood
  - model adequacy
- applications, including to regression and factorial experiments
- demonstrations and practical work using **R**

# Some data sets

The kyphosis data frame has 81 rows and 4 columns, data on children who have had corrective spinal surgery
Kyphosis - a factor with levels absent present indicating if a kyphosis (a type of deformation) was present after the operation.
Age - in months
Number - the number of vertebrae involved
Start - the number of the first (topmost) vertebra operated on.

```
> head(kyphosis)
   Kyphosis Age Number Start
1    absent  71      3     5
2    absent 158      3    14
3   present 128      4     5
4    absent   2      5     1
5    absent   1      4    15
6    absent   1      2    16
```

# Some data sets

The oxygen data frame has 24 rows and 2 columns, data on maximal oxygen uptake in active military personnel during treadmill running
For each of 24 males, these variables were measured: VO2max - the maximum volume of oxygen uptake in the blood.
time - time taken to run 2 miles (in seconds)

```
> head(oxygen)
  VO2max time
1   42.3  918
2   53.1  805
3   42.1  892
4   50.1  962
5   42.5  968
6   42.5  907
```

# Some data sets

The rubber data frame has 30 rows and 3 columns. Physical properties of 30 rubber samples. Abrasion is expensive to determine, so there is interest in predicting it from Hardness and Tensile.

Abrasion : abrasion loss (g/hp-hour) Hardness : hardness (deg shore) Tensile : tensile strength (kg/sqcm)

```
> head(rubber)
  Abrasion Hardness Tensile
1      372       45     162
2      206       55     233
3      175       61     232
4      154       66     231
5      136       71     231
6      112       71     237
```

# Some data sets

The cabbages data frame shows the results of a field trial on the cultivation of cabbages, giving the yields in each plot.

**Cabbage yields**

| | | | | |
|---|---|---|---|---|
| 396 | 559 | 417 | 465 | 312 |
| 399 | 498 | 442 | 590 | 438 |
| 610 | 596 | 538 | 568 | 346 |

# Some data sets

The potatoes data frame shows the results of a field trial on the cultivation of potatoes, giving the yields in each plot and the levels of nitrate and phosphate fertilizers applied to the plot.

**Potato yields**

| | | | | | |
|---|---|---|---|---|---|
| 416<br>high<br>low | 282<br>low<br>low | 646<br>high<br>high | 384<br>low<br>mid | 466<br>low<br>high | 420<br>high<br>mid |
| 504<br>low<br>high | 571<br>high<br>high | 334<br>high<br>low | 323<br>low<br>low | 594<br>high<br>mid | 326<br>low<br>mid |
| 390<br>low<br>low | 488<br>high<br>mid | 422<br>low<br>mid | 439<br>low<br>high | 617<br>high<br>high | 366<br>high<br>low |
| 652<br>high<br>high | 415<br>high<br>low | 483<br>low<br>high | 505<br>high<br>mid | 411<br>low<br>mid | 259<br>low<br>low |
| 527<br>low<br>mid | 475<br>low<br>high | 481<br>high<br>mid | 448<br>high<br>low | 432<br>low<br>low | 505<br>high<br>high |
| 633<br>high<br>mid | 489<br>low<br>mid | 384<br>low<br>low | 620<br>high<br>high | 452<br>high<br>low | 500<br>low<br>high |

Key:

| |
|---|
| Yield<br>nit<br>phos |

# Motivation: data sets and basic ideas

- structure and relationships
- response and explanatory variables
- quantitative and qualitative (categorical, factor) variables
- statistical modelling
- experiment and observation
- causation
- estimation, confidence intervals, testing
- prediction

Linear models play a central role in theoretical and applied statistics

- in practice – a major part of the basic toolkit
- pedagogically – a pattern for other techniques

# 4. Model formulation

Linear models can be formulated or specified in various ways. It is important to be able to translate models fluently from one specification to another.

Responses are known linear functions of unknown parameters, plus an error term. In matrix/vector notation:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y}$ is $n \times 1$, and $X$ is $n \times p$, where $p < n$.

The *response vector* $\boldsymbol{Y}$ is assumed known (observed); the *model* or *design* matrix $X$ is also known, comprising observed explanatory variables or experimental settings. The parameter vector $\boldsymbol{\beta}$ is the focus of our interest, whether it is to be estimated, or some other inference carried out on it.

$$\boldsymbol{Y} = X\beta + \varepsilon,$$

or, spelling it out:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Quite often, the model includes a constant or intercept term, which we will may refer to as column 0:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Here, $p$ is $k+1$.

Matrix notation is convenient for developing the general theory, but specific models are usually specified either in ordinary *algebraic notation* or in the mnemonic *model expression* notation, devised by Wilkinson and Rogers, that is used in many computer systems, including **R**.

Let us illustrate this with examples.

# a. Simple linear regression

Algebraic notation:

$$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, \ldots, n$$

Matrix formulation: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as usual, with

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Model expression: `Y ~ x` (or, more commonly, we would use self-explanatory names for the variables, e.g. `cholesterol ~ age`). Note that the intercept is not mentioned, but is included by default.

# b. Multiple linear regression

Algebraic notation:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, 2, \ldots, n$$

Matrix formulation:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Model expression: (e.g.)

```
Abrasion ~ Hardness+Tensile
```

# c. Linear regression without an intercept

Algebraic notation:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, 2, \ldots, n$$

Matrix formulation:

$$X = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Model expression: (e.g.)

```
Abrasion ~ Hardness+Tensile-1
```

# d. Linear regression with functions and combinations of explanatory variables

Algebraic notation:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 \log x_{i2} + \beta_3 x_{i1} \sin(x_{i2}) + \epsilon_i$$

Matrix formulation:

$$X = \begin{pmatrix} 1 & x_{11} & \log(x_{12}) & x_{11}\sin(x_{12}) \\ \vdots & & & \vdots \\ 1 & x_{n1} & \log(x_{n2}) & x_{n1}\sin(x_{n2}) \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Model expression: (e.g.)

```
Abrasion ~ Hardness+log(Tensile)
     +Hardness:sin(Tensile)
```

# e. Factor variables

Algebraic notation:

$$Y_i = \alpha_{s_i} + \epsilon_i, i = 1, 2, \dots, n$$

where $s_i \in \{1, 2, \dots, k\}$ is an observed factor (that is, a qualitative variable). Matrix formulation:

$$X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$$

The columns of $X$ are often called *dummy variables*; they are not observed numbers, but indicators of which component of $\boldsymbol{\beta}$ enters into the formula for that observation.

Model expression: (e.g.)

```
lifetime ∼ make
```

Such models are also often expressed in a double-subscript notation:

$$Y_{ij} = \alpha_i + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$$

where $n = \sum_{i=1}^{k} n_k$. This is the same model in a different notation; $\boldsymbol{Y}$ is not a matrix, but a vector: $\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{kn_k} \end{pmatrix}$

# f. Two factors: row+column model

Algebraic notation (with double subscripts):

$$Y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}, i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$$

Matrix formulation:

$$X = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} & I \\ \vdots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} & I \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \\ \beta_1 \\ \vdots \\ \beta_c \end{pmatrix}$$

In $X$, each $\mathbf{1}$ and $\mathbf{0}$ is a $c$-vector of 1's and 0's, respectively, and each $I$ is the $c \times c$ identity matrix. There are $r$ rows of these blocks, so $X$ is $rc \times (r + c)$.
Note that we return to this set-up later, and use a slightly different notation.
Model expression: (e.g.)

```
Yield ~  Site+Variety
```

# g. Regression, with factor variables too

Algebraic notation:

$$Y_i = \alpha_{s_i} + \beta x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where $s_i \in \{1, 2, \ldots, k\}$ is an observed factor, and $x_i$ an ordinary numerical variable. Matrix formulation:

$$X = \begin{pmatrix} 1 & 0 & \cdots & 0 & x_1 \\ 1 & 0 & \cdots & 0 & x_2 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \\ \beta \end{pmatrix}$$

(assuming $s_1 = s_2 = 1$ and $s_n = k$).
Model expression: (e.g.)

```
lifetime ~ make+speed
```

In double-subscript notation:

$$Y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$$

Note that this specifies several *parallel* regression lines.

# h. Regression, with factor variables and interaction

Algebraic notation:

$$Y_i = \alpha_{s_i} + \beta_{s_i} x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where $s_i \in \{1, 2, \ldots, k\}$ is an observed factor.
Model expression: (e.g.)
`lifetime ~ make*speed` or, equivalently,
`lifetime ~ make+speed+make:speed`
In double-subscript notation:

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$$

Note that this specifies several *separate* regression lines: they need no longer be parallel.

# Alternative parameterisations

There are always several ways to parameterise a model, and when interpreting parameters or their estimates, it is important to bear the parameterisation in mind.
For example

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \epsilon_i, i = 1, 2, \ldots, n \\ \text{and} \quad Y_i &= \alpha^\star + \beta(x_i - \overline{x}) + \epsilon_i, i = 1, 2, \ldots, n \end{aligned}$$

are identical models, where $\alpha^\star = \alpha + \beta\overline{x}$. While $\beta$ is the gradient in both models, the "intercepts" $\alpha$ and $\alpha^\star$ are different numbers with different meanings.
This issue is especially important with factor variables. For example in case (e) above, with two levels of the factor $s_i$ as in the `lathe` example, we might have chosen $\alpha_1 = \mu$ and $\alpha_2 = \mu + \delta$, so that $\delta$ is the *difference* in mean lifetime between the two makes of lathe. The $X$ matrix would then be of the form

$$X = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} \end{pmatrix} \quad \text{instead of} \quad \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$$

Mathematically, two $X$ matrices represent the same model if and only if they have the same *column spaces*.

# Summary

- Note the general patterns, so that you can put richer models together using these ideas in a modular way.
- A numerical explanatory variable contributes one column to $X$; a factor contributes as many columns as there are *levels* of the factor.
- We use the $+$ symbol to assemble sets of columns together, corresponding to *adding* in successive terms in the model.
- Learn the interpretation of *interaction* (symbolised by `:`) between two factors, and between a factor and a numerical variable.
- `A*B` is short for `A+B+A:B`
- The individual components of $\beta$ might correspond to other greek letters in the algebraic specification of the model.
- The individual components of $\mathbf{Y}$ and $\varepsilon$ might have multiple subscripts in the algebraic specification of the model.

# 5. Least squares estimation

Fitting a linear model means estimating the regression coefficient parameter $\boldsymbol{\beta}$ – we usually do this using the principle of least squares. An advantage of this principle is that it makes sense without having to assume a statistical model for the errors $\boldsymbol{\varepsilon}$.

The idea is to choose that value of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, such that the *residual sum of squares* $S(\boldsymbol{\beta})$ is minimised, where

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = ||\boldsymbol{\varepsilon}||^2$$

where

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - X\boldsymbol{\beta}$$

We can find a compact general expression for the solution to this minimisation problem, if we make a simplifying assumption.

Until further notice we assume:

The matrix $X$ is of full rank. This is equivalent to making any of these assertions (recalling that $p < n$):

- The rank of $X$ is $p$
- The columns of $X$ are linearly independent
- $X^T X$ is non-singular

(Note that this is a sensible assumption, since if it was not true, $X\beta = X\beta^\star$ would not imply $\beta = \beta^\star$, so that you could not expect to choose between $\beta$ and $\beta^\star$ using data $\boldsymbol{Y} = X\beta + \boldsymbol{\varepsilon}$. In this case, we say $\beta$ is *not identifiable*.)

Which of the cases (a) to (h) in Section 1 satisfy this assumption?

Let $\widehat{\beta}$ satisfy

$$(X^T X)\widehat{\beta} = X^T \boldsymbol{Y}. \tag{1}$$

We shall prove that any such $\widehat{\beta}$ minimises $S(\beta)$.
Note that $S(\beta) = (\boldsymbol{Y} - X\beta)^T(\boldsymbol{Y} - X\beta)$. Let $\boldsymbol{\delta}$ be any $p$-vector. Then

$$S(\widehat{\beta} + \boldsymbol{\delta}) - S(\widehat{\beta})$$

$$= (\boldsymbol{Y} - X\widehat{\beta} - X\boldsymbol{\delta})^T(\boldsymbol{Y} - X\widehat{\beta} - X\boldsymbol{\delta}) - (\boldsymbol{Y} - X\widehat{\beta})^T(\boldsymbol{Y} - X\widehat{\beta})$$

$$= -2(X\boldsymbol{\delta})^T(\boldsymbol{Y} - X\widehat{\beta}) + (X\boldsymbol{\delta})^T(X\boldsymbol{\delta})$$

Simplifying, the first term on the right is $-2\boldsymbol{\delta}^T(X^T\boldsymbol{Y} - (X^T X)\widehat{\beta}) = 0$ by assumption (1), for all $\boldsymbol{\delta}$ and $\boldsymbol{Y}$.

The second term on the right is the sum of squares of the elements of $X\delta$, so is non-negative, and is 0 if and only if $X\delta = 0$. Since $X$ is full rank, this is true if and only if $\delta = 0$.

Thus, $S(\widehat{\beta} + \delta) - S(\widehat{\beta}) \geq 0$, with equality if and only if $\delta = 0$. So any $\widehat{\beta}$ satisfying (1) minimises $S(\beta)$.

Since $X^T X$ is nonsingular, the only solution is $\widehat{\beta} = (X^T X)^{-1} X^T \boldsymbol{Y}$; this is therefore the unique least squares estimator.

# Fitted values and residuals

Having obtained the estimates $\widehat{\beta}$, the predicted or *fitted values* of the response variable are obtained by substitution:

$$\widehat{\boldsymbol{Y}} = X\widehat{\beta} = X(X^TX)^{-1}X^T\boldsymbol{Y} = H\boldsymbol{Y}$$

where $H = X(X^TX)^{-1}X^T$ is called the hat matrix (it 'puts the hat' on $\boldsymbol{Y}$). Similarly, the vector of *residuals* is the difference

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I - X(X^TX)^{-1}X^T)\boldsymbol{Y} = (I - H)\boldsymbol{Y}$$

where $I = I_n$ is the order $n$ identity matrix. Note that both the fitted values and the residuals are, like the least squares estimates $\widehat{\beta}$, linear functions of $\boldsymbol{Y}$.

The *residual sum of squares* is

$$\boldsymbol{e}^T\boldsymbol{e} = S(\widehat{\beta}) = \boldsymbol{Y}^T(I-H)^T(I-H)\boldsymbol{Y} = \boldsymbol{Y}^T(I-H)\boldsymbol{Y}$$

since, as you can show easily:

- $H$ is symmetric: $H^T = H$
- $H$ is idempotent: $H^2 = H$, and so:
- $(I-H)^T(I-H) = (I-H)$

# Fitting linear models in R

The **R** command for fitting a linear model is `lm()`; the only compulsory argument is the *formula* of the model to be fitted – in the model expression syntax we saw in Section 1. For example,

```
lm(Cholesterol∼Age)
```

The variables used in the formula may be either (i) in the current workspace as ordinary variables, (ii) in a data frame that has been previously attached using the command, e.g. `attach(lipid)`, or (iii) in a data frame specified as the 2nd argument of `lm()`, e.g.

```
lm(Cholesterol∼Age,lipid)
```

It produces brief output, the least squares estimates.

More comprehensive output is obtained by assigning the output of `lm()` to a variable, with a name of your choosing, e.g.

```
fit1<-lm(Cholesterol~Age)
```

and then processing the result.
The output from `lm()` is a list with 12 named components, e.g. `fit1$residuals`; you can see all the names with, e.g.,

```
names(fit1)
```

You can look at the values of these components, as usual, by typing the name, e.g. `fit1$coef` gives the least squares estimates.

# 6. Statistical performance

In this section, we start to make statistical assumptions about our model, but only about means and variances, not full probability distributions. Remarkably, this is enough to demonstrate a particular kind of optimality of least squares estimators, in the form of a famous result known as the Gauss-Markov theorem.

# Mean and variance assumptions

In our linear model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

from now on, we assume that

- $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ (or, equivalently, $E(\boldsymbol{Y}) = X\boldsymbol{\beta}$): that is, the observations are *unbiased*;
- $\text{var}(\epsilon_i) = \sigma^2$ for all $i$ (or, equivalently, $\text{var}(Y_i) = \sigma^2$ for all $i$): that is, the observations have *equal variance*;
- $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$ (or, equivalently, $\text{cov}(Y_i, Y_j) = 0$ for all $i \neq j$): that is, the observations are *uncorrelated*.

The 2nd and 3rd items are the same as saying that $\text{var}(\boldsymbol{\varepsilon}) = \text{var}(\boldsymbol{Y}) = \sigma^2 I_n$.

# Mean and variance of the least squares estimator

We find that

$$E(\widehat{\beta}) = \beta \quad \text{and} \quad \text{var}(\widehat{\beta}) = \sigma^2 (X^T X)^{-1}$$

*Proof*
We will be using the results that

for any random $n$-vector $\boldsymbol{Y}$, and any constant $m \times n$ matrix $A$, $E(A\boldsymbol{Y}) = AE(\boldsymbol{Y})$ and $\text{var}(A\boldsymbol{Y}) = A\text{var}(\boldsymbol{Y})A^T$.

Recall that under the full-rank assumption, $\widehat{\beta} = (X^T X)^{-1} X^T \boldsymbol{Y}$. Then

$$E(\widehat{\beta}) = E((X^T X)^{-1} X^T \boldsymbol{Y}) = (X^T X)^{-1} X^T E(\boldsymbol{Y})$$

$$= (X^T X)^{-1} X^T X \beta = \beta$$

and

$$\mathrm{var}(\widehat{\beta}) = \mathrm{var}((X^T X)^{-1} X^T \boldsymbol{Y})$$

$$= (X^T X)^{-1} X^T \mathrm{var}(\boldsymbol{Y}) [(X^T X)^{-1} X^T]^T$$

$$= (X^T X)^{-1} X^T [\sigma^2 I_n] X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

These results allow us to say that least squares estimates are unbiased, and to write down their standard errors

$$\sqrt{\text{var}(\widehat{\beta}_j)} = \sigma^2[(X^TX)^{-1}]_{jj}.$$

Confidence intervals, etc., based on this will be in the next section. But even without the further assumptions made there, we can claim a remarkable optimality for least squares estimation; the variance of the estimator is the *smallest possible* (among linear unbiased estimators).

**Theorem (Gauss–Markov)**

*Suppose that $E(\boldsymbol{Y}) = X\beta$ and $var(\boldsymbol{Y}) = \sigma^2 I_n$. Let $\boldsymbol{c}$ be a fixed p-vector. Then $\boldsymbol{c}^T\widehat{\beta}$ is an unbiased estimator of $\boldsymbol{c}^T\beta$, and has variance smaller than any other estimator that is linear (in $\boldsymbol{Y}$) and unbiased.*

*Proof*

$$E(\boldsymbol{c}^T\widehat{\beta}) = \boldsymbol{c}^T E(\widehat{\beta}) = \boldsymbol{c}^T\beta.$$

Also, $\boldsymbol{c}^T\widehat{\beta} = \boldsymbol{c}^T(X^TX)^{-1}X^T\boldsymbol{Y}$ is clearly linear in $\boldsymbol{Y}$. Let $\boldsymbol{\lambda}^T\boldsymbol{Y}$ be any other linear unbiased estimator. Then

$$\boldsymbol{c}^T\beta = E(\boldsymbol{\lambda}^T\boldsymbol{Y}) = \boldsymbol{\lambda}^T E(\boldsymbol{Y}) = \boldsymbol{\lambda}^T X\beta$$

holds for all $\boldsymbol{\beta}$, that is, $\boldsymbol{c}^T = \boldsymbol{\lambda}^T X$.

So

$$\mathrm{var}(\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}) = \boldsymbol{c}^T\sigma^2(X^TX)^{-1}\boldsymbol{c} = \boldsymbol{\lambda}^TX\sigma^2(X^TX)^{-1}(\boldsymbol{\lambda}^TX)^T$$

$$= \sigma^2\boldsymbol{\lambda}^TX(X^TX)^{-1}X^T\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}^TH\boldsymbol{\lambda}$$

Meanwhile,

$$\mathrm{var}(\boldsymbol{\lambda}^T\boldsymbol{Y}) = \boldsymbol{\lambda}^T\sigma^2I_n\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}^T\boldsymbol{\lambda}$$

so

$$\mathrm{var}(\boldsymbol{\lambda}^T\boldsymbol{Y}) - \mathrm{var}(\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\lambda}^T\boldsymbol{\lambda} - \sigma^2\boldsymbol{\lambda}^TH\boldsymbol{\lambda}$$

$$= \sigma^2\boldsymbol{\lambda}^T(I_n - H)\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}^T(I_n - H)^T(I_n - H)\boldsymbol{\lambda} \geq 0$$

as required.

# Estimating $\sigma^2$

The least squares principle does not tell itself us how to estimate $\sigma^2$.
However, we do now have a basis for doing so. Since $\sigma^2 = \text{var}(Y_i)$ and $E(Y_i) = (X\beta)_i$, we would expect the average of the squares of the residuals $Y_i - (X\widehat{\beta})_i$ to be about $\sigma^2$.
In fact (from page 19) the residual sum of squares is $\boldsymbol{e}^T\boldsymbol{e} = S(\widehat{\beta}) = \boldsymbol{Y}^T(I_n - H)\boldsymbol{Y}$, and using the result that

For any random vector $\boldsymbol{Y}$ with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{var}(\boldsymbol{Y}) = V$, and any constant matrix $A$, $E(\boldsymbol{Y}^TA\boldsymbol{Y}) = \boldsymbol{\mu}^TA\boldsymbol{\mu} + \text{tr}(AV)$.

we have

$$E(\boldsymbol{e}^T\boldsymbol{e}) = E(\boldsymbol{Y}^T(I_n - H)\boldsymbol{Y})$$

$$= (X\beta)^T(I_n - H)(X\beta) + \text{tr}((I_n - H)\sigma^2 I_n) = \sigma^2(n - \text{tr}(H))$$

since $HX = X$ (check!).

# Estimating $\sigma^2$ [2]

But

$$\mathrm{tr}(H) = \mathrm{tr}(X(X^TX)^{-1}X^T)$$

$$= \mathrm{tr}(X^TX(X^TX)^{-1}) = \mathrm{tr}(I_p) = p,$$

using the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for all conformable matrices $A$, $B$.
Thus,

$$\widehat{\sigma}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p}$$

("RSS $\div$ df") is an unbiased estimator of $\sigma^2$.

# Standardised residuals

Although the errors $\epsilon_i$ are assumed to have equal variance $\sigma^2$, their estimates, the residuals $e_i$, do not. In fact,

$$\text{var}(e_i) = [\text{var}(\boldsymbol{e})]_{ii} = [\text{var}((I - H)\boldsymbol{Y})]_{ii}$$

$$= [\sigma^2(I - H)]_{ii} = \sigma^2(1 - h_{ii})$$

where $h_{ii}$ are the diagonal elements of $H$.
Therefore we define *standardised residuals* as
$e'_i = e_i/(\widehat{\sigma}\sqrt{1 - h_{ii}}), i = 1, 2, \ldots, n$. These have (approximately) equal variances.

# Further results from `lm()`

A more complete printed summary can be obtained by typing, e.g.,

```
summary(fit1)
```

This includes estimates and their standard errors, and some statistics about residuals and about the fit.
Four diagnostic plots are produced if you type, e.g.,

```
plot(fit1)
```

(If you have first typed `par(mfrow=c(2,2))` they will be displayed as a $2 \times 2$ array.)

# Diagnostic plots

The four plots are

1. Fitted values vs. residuals: a scatter plot of $(\widehat{Y}_i, e_i)$
   (a pattern indicates that there is systematic under-fitting: do you need to fit other explanatory variables?)
2. Normal Q-Q plot: a Q-Q plot of the standardised residuals $e'_i$
   (departures from a straight line suggest that the errors are not normally distributed)
3. Scale-Location plot: a scatter plot of $(\widehat{Y}_i, \sqrt{|e'_i|})$
   (a more sensitive version of the fitted value/residual plot: does the variance of the errors vary?)
4. Cook's distance plot: a plot of $e'^2_i h_{ii}/(p(1 - h_{ii})$ against $i$
   (Cook's distance is a measure of how much the overall fit would change if observation $i$ was deleted)

# Prediction

We often fit a linear model in order to make predictions of the response variable for various future choices of the explanatory variable(s). The function `predict()` is provided for this.
For example,

```
predict(fit1,data.frame(Age=c(23,27)))
```

computes the expected values of `Cholesterol` for `Age` equal to 23 and 27.

# 7. Normal theory assumptions

Now for the first time, we make assumptions about the probability distribution of our responses. We assume more, and we get more - we can derive inferential procedures like confidence intervals for parameters, make probabilistic predictions about future observations, and test hypotheses about parameter values and about model adequacy.
We will also find an intimate connection between least squares and maximum likelihood.

# Assumption

In addition to the assumptions of Section 3, we now assume that the $\{Y_i\}$ are independently normally distributed. That is, $Y_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)$, independently, or in brief, $\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$.
(Here $\boldsymbol{x}_i$ is the $i^{\text{th}}$ row of $X$; so $\boldsymbol{x}_i^T \boldsymbol{\beta} = (X\boldsymbol{\beta})_i$.)

# Least squares and maximum likelihood

Since the observations are independent, the likelihood is just the product of their density functions, so

$$L = L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{x}_i^T\beta)^2\right\}$$

Thus the log-likelihood is

$$\ell = \ell(\beta, \sigma^2) = -n\log\sqrt{2\pi} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\beta)^2$$

$$= -n\log\sqrt{2\pi} - n\log\sigma - \frac{1}{2\sigma^2}S(\beta),$$

where

$$S(\beta) = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{Y} - X\beta)^T(\boldsymbol{Y} - X\beta)$$

is the usual sum of squares.

Hence for any $\sigma$, maximising the likelihood corresponds exactly to minimising the sum of squares of the residuals. That is, for $\boldsymbol{\beta}$, least squares estimation and maximum likelihood estimation is the same thing. (Note that the normal distribution assumption is essential for this conclusion.)

Since the least squares estimates do not involve $\sigma^2$, you get the same estimators on *simultaneously* maximising over $\beta$ and $\sigma^2$.
This connection provides a powerful additional justification for using least squares estimators.
Differentiating the log-likelihood with respect to $\sigma$, and setting to zero, we get

$$-\frac{n}{\sigma} + \frac{S(\beta)}{\sigma^3} = 0$$

so we immediately obtain the maximum likelihood estimator of $\sigma^2$ as $S(\widehat{\beta})/n$.
Note that this is *different* from the least squares estimator $\widehat{\sigma}^2$, which has the divisor $(n-p)$; in practice we always use the latter, the unbiased estimator.

# Joint distribution of $\widehat{\beta}$ and $S(\widehat{\beta})$

Various inferential procedures can be derived from the following result:
If $\boldsymbol{Y} \sim N_n(X\beta, \sigma^2 I)$ then

(a) $\widehat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$

(b) $S(\widehat{\beta})/\sigma^2 \sim \chi^2_{n-p}$

(c) $\widehat{\beta}$ and $S(\widehat{\beta})$ are independent.

Note that we already knew the mean and variance in (a) and the mean in (b) – they did not require normality.

# Corollary

For any fixed $p$-vector $\boldsymbol{c}$, we have

$$\frac{\boldsymbol{c}^T\widehat{\beta} - \boldsymbol{c}^T\beta}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

*Proof*
From (a), $\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{c}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c})$, so

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}}} \sim N(0,1) \tag{2}$$

From (b), $\widehat{\sigma}^2/\sigma^2 = S(\widehat{\boldsymbol{\beta}})/((n-p)\sigma^2) \sim \chi^2_{n-p}/(n-p)$, and this is independent of (2) by (c), so we have our result, by the definition of the $t$ distribution:

$$\text{“} \quad t_\nu = N(0,1)/\sqrt{\chi^2_\nu/\nu} \quad \text{”}$$

Thus, for example, a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{c}^T\boldsymbol{\beta}$ is given by

$$[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - t_{n-p}(\alpha/2)\widehat{\sigma}\omega, \boldsymbol{c}^T\widehat{\boldsymbol{\beta}} + t_{n-p}(\alpha/2)\widehat{\sigma}\omega]$$

where $\omega = \sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}}$.

# Example 1: confidence interval

A $100(1-\alpha)\%$ confidence interval for $\beta_j$is given by

$$[\widehat{\beta}_j - t_{n-p}(\alpha/2)\widehat{\sigma}\omega, \widehat{\beta}_j + t_{n-p}(\alpha/2)\widehat{\sigma}\omega]$$

where now $\omega = \sqrt{[(X^TX)^{-1}]_{jj}}$.
You can reject the hypothesis that $\boldsymbol{c}^T\boldsymbol{\beta} = \boldsymbol{c}^T\boldsymbol{\beta}_0$ at level $\alpha$, against a two-sided alternative, if

$$\left|\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}_0}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}}}\right| > t_{n-p}(\alpha/2)$$

A future observation with explanatory variables $\boldsymbol{x}_\star$will be $\boldsymbol{Y}_\star = \boldsymbol{x}_\star^T\beta + \varepsilon_\star$; this has least squares estimate $\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}}$. The error has variance $\sigma^2\boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star + \sigma^2$. A $100(1-\alpha)\%$ prediction interval for $\boldsymbol{Y}_\star$ is

$$[\boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}} - t_{n-p}(\alpha/2)\widehat{\sigma}\omega, \boldsymbol{x}_\star^T\widehat{\boldsymbol{\beta}} - t_{n-p}(\alpha/2)\widehat{\sigma}\omega]$$

where this time $\omega = \sqrt{1 + \boldsymbol{x}_\star^T(X^TX)^{-1}\boldsymbol{x}_\star}$.

# 8. Model choice in linear models

We have so far regarded the matrix $X$, containing numerical explanatory variables, and 0/1 indicators for factor levels, as fixed and given. In practice, very often, choice of which explanatory variables and factors to include is at the discretion of the analyst. As we have seen, it is easy enough in a system like **R** to make several different choices, and fit them all. But what criteria should be used to choose between these models?
We wish to avoid

- missing out variables that are important – that would incur bias in estimation and prediction
- including variables that have no effect – that is a waste of expense, and would lead to inflated estimates of variance of estimates and predictions

The main thing we need is a formal mechanism for determining whether an individual variable, or group of variables, can be dropped from a linear model, without an undue effect on the performance of the model in terms of explaing the variation in $Y$. This will allow us to make *pairwise* comparisons between models.

We already have a method for testing whether a single variable (component of $\boldsymbol{x}$) needs to be included – see Example 2 on page 37 . If we set $\boldsymbol{c}$ to be the $j^{\text{th}}$ unit vector, so $\boldsymbol{c}^T\boldsymbol{\beta} = \beta_j$, and suppose $(\beta_0)_j = 0$ then we see that you can reject the hypothesis that $\beta_j = 0$at level $\alpha$, against a two-sided alternative, if

$$\left| \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{[(X^TX)^{-1}]_{jj}}} \right| > t_{n-p}(\alpha/2)$$

However, this procedure doesn't cover the case where the possible exclusion of several components of $\boldsymbol{x}$ is being considered, since the test for each component assumes the inclusion of all other components.

# Analysis of Variance

This is a general term for procedures that give the generalisation of the $t$test that we require.

The basic idea is to *decompose* the variability, measured by sums of squares, among components in $\boldsymbol{Y}$ into terms attributable to different components of $\boldsymbol{x}$.

Tests are based on ratios between these sums of squares.

The basic decomposition is obtained from results on page 19:

$$H^T H = H^2 = H \text{ and } (I-H)^T(I-H) = (I-H)^2 = I-H$$

thus:

$$I = H + (I-H) = H^T H + (I-H)^T(I-H)$$

Post- and pre-multiplying by $\boldsymbol{Y}$:

$$\boldsymbol{Y}^T\boldsymbol{Y} = \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}} + (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$$

since $\widehat{\boldsymbol{Y}} = H\boldsymbol{Y}$. In symbols,

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{E}}$$

where $SS_{\mathrm{T}}$, $SS_{\mathrm{R}}$ and $SS_{\mathrm{E}}$ are called the total, regression and residual (or error) sums of squares, respectively. (This is really Pythagoras' theorem!) Because of the special properties of $H$, each of these terms can be expressed in many equivalent ways.
What we have done is decompose the variation in $\boldsymbol{Y}$into a term explained by the regression, and an unexplained, or residual, term.

If $SS_{\mathrm{R}}$ is large relative to $SS_{\mathrm{E}}$, we intuitively conclude that the regression is doing a good job; now we develop a formal test to assess this.
We know (slides 27/8) that $E(SS_{\mathrm{E}}) = (n-p)\sigma^2$. Meanwhile,

$$E(SS_{\mathrm{R}}) = E(\widehat{\boldsymbol{Y}}^T \widehat{\boldsymbol{Y}}) = E(\widehat{\boldsymbol{\beta}}^T X^T X \widehat{\boldsymbol{\beta}})$$

$$= \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} + \mathrm{tr}[X^T X \mathrm{var}(\widehat{\boldsymbol{\beta}})]$$

$$= \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} + \mathrm{tr}[X^T X \sigma^2 (X^T X)^{-1}] = \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} + p\sigma^2$$

So if $\boldsymbol{\beta} = \mathbf{0}$, $SS_{\mathrm{R}}/p$ is another unbiased estimator of $\sigma^2$, and

$$F = \frac{SS_{\mathrm{R}}/p}{SS_{\mathrm{E}}/(n-p)}$$

should be close to 1; if $\boldsymbol{\beta} \neq \mathbf{0}$ it will tend to be larger.

# The $F$ test

If we assume, as usual, that $\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$ and that $X$ has full rank $p$, then

(a) if $\boldsymbol{\beta} = \mathbf{0}$, $SS_{\mathrm{R}} \sim \sigma^2 \chi^2_p$

(b) always, $SS_{\mathrm{E}} \sim \sigma^2 \chi^2_{n-p}$

(c) $SS_{\mathrm{E}}$ and $SS_{\mathrm{R}}$ are independent.

It follows by definition of the $F$ distribution that if $\boldsymbol{\beta} = \mathbf{0}$,

$$F \sim F_{p,n-p}$$

*(Actually, regarding (a), in general $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \sigma^2 \chi^2_p$, and if $\boldsymbol{\beta} \neq \mathbf{0}$ then $SS_{\mathrm{R}}/\sigma^2$ has what we call a non-central $\chi^2$ distribution.)*

The resulting procedure of calculating the sums of squares and other terms to perform this F test is usually summarised as an ANOVA table:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression on $\beta$ | $SS_{\mathrm{R}}$ | $p$ | $SS_{\mathrm{R}}/p$ | $F$ |
| Residual | $SS_{\mathrm{E}}$ | $n-p$ | $SS_{\mathrm{E}}/(n-p)$ | |
| Total, uncorrected | $SS_{\mathrm{T}}$ | $n$ | | |

# Correcting for the mean

More often than not, variation in $Y$ is of interest measured from the mean $\overline{Y}$ of $Y$; $SS_{\mathrm{R}}$ and $SS_{\mathrm{T}}$ are then modified accordingly:

$$SS_{\mathrm{R}}^{\star} = SS_{\mathrm{R}} - n\overline{Y}^2 \qquad SS_{\mathrm{T}}^{\star} = SS_{\mathrm{T}} - n\overline{Y}^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

The ANOVA table is modified to:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression on $\beta$ | $SS_{\mathrm{R}}^{\star}$ | $p-1$ | $SS_{\mathrm{R}}^{\star}/(p-1)$ | $F$ |
| Residual | $SS_{\mathrm{E}}$ | $n-p$ | $SS_{\mathrm{E}}/(n-p)$ | |
| Total, corrected | $SS_{\mathrm{T}}^{\star}$ | $n-1$ | | |

where now $F \sim F_{p-1,n-p}$

This $F$ ratio is appropriate for testing the hypothesis that $\beta_2 = \beta_3 = \cdots = \beta_p = 0$ in the model

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, i = 1, 2, \ldots, n$$

where $x_{i1} = 1$ for all $i$, so that $\beta_1$ is the intercept. We reject that hypothesis at level $\alpha$ if $F > F_{p-1,n-p}(\alpha)$.

# Significance of subsets of variables

Having fitted, say, $x_1, x_2, \ldots, x_p$, were $x_{q+1}, \ldots, x_p$ really necessary? (In practice, we may re-order variables before posing this question. We suppose that $x_1 \equiv 1$: the model always includes an intercept.)
We answer this question by comparing the fits of two models using a significance test. The models are:
The **Full** model: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
The **Reduced** model: $\boldsymbol{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$
where we have partitioned $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ and $X = [X_1|X_2]$ where $\boldsymbol{\beta}_1$ is $q \times 1$, $\boldsymbol{\beta}_2$ is $(p-q) \times 1$, $X_1$ is $n \times q$ and $X_2$ is $n \times (p-q)$. [NB – be careful, $\boldsymbol{\beta}_1$ here is still a vector, the first $p$ components of $\boldsymbol{\beta}$, not just the first component.]
The question: are $x_{q+1}, \ldots, x_p$ really necessary? becomes now: can we accept the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$?

We fit both models by least squares, and obtain least squares estimates $\widehat{\beta}_{\mathrm{F}}$and $\widehat{\beta}_{1\mathrm{R}}$, and residual sums of squares $SS_{\mathrm{EF}} = S(\widehat{\beta}_{\mathrm{F}})$ and $SS_{\mathrm{ER}} = S(\widehat{\beta}_{1\mathrm{R}})$. Obviously $SS_{\mathrm{ER}} - SS_{\mathrm{EF}} >= 0$ – how much larger says how strongly it was worth including the explanatory variables in $X_2$. This is the basis of the test – we need to find the distribution of this difference under $H_0$.

Suppose the full model is true, then

(a) if $H_0$ is true, $SS_{\mathrm{ER}} - SS_{\mathrm{EF}} \sim \sigma^2 \chi^2_{p-q}$

(b) always, $SS_{\mathrm{EF}} \sim \sigma^2 \chi^2_{n-p}$

(c) $SS_{\mathrm{EF}}$ and $SS_{\mathrm{ER}} - SS_{\mathrm{EF}}$ are independent.

Them under $H_0$,

$$F = \frac{SS_{\mathrm{ER}} - SS_{\mathrm{EF}}}{p-q} \div \frac{SS_{\mathrm{EF}}}{n-p} \sim F_{p-q,n-p}$$

and it otherwise tends to be bigger, so we reject $H_0$ at the significance level $\alpha$ if $F > F_{p-q,n-p}(\alpha)$.

The main part of the ANOVA table becomes:

| Source | SS | df |
|---|---|---|
| Regression on $\boldsymbol{\beta}_1$ | $SS^\star_{\mathrm{RR}}$ | $q-1$ |
| Due to $\boldsymbol{\beta}_2$ after $\boldsymbol{\beta}_1$ | $SS^\star_{\mathrm{RF}} - SS^\star_{\mathrm{RR}}$ | $p-q$ |
| Regression on $\boldsymbol{\beta}$ | $SS^\star_{\mathrm{RF}}$ | $p-1$ |
| Residual | $SS_{\mathrm{EF}}$ | $n-p$ |
| Total, corrected | $SS^\star_{\mathrm{T}}$ | $n-1$ |

using the fact that $SS^\star_{\mathrm{T}} = SS^\star_{\mathrm{RF}} + SS_{\mathrm{EF}} = SS^\star_{\mathrm{RR}} + SS_{\mathrm{ER}}$, so that $SS_{\mathrm{ER}} - SS_{\mathrm{EF}} = SS^\star_{\mathrm{RF}} - SS^\star_{\mathrm{RR}}$.

If we need to do the computations by hand, usually it is easiest to compute $SS^\star_{\mathrm{T}}$, $SS^\star_{\mathrm{RR}}$ and $SS^\star_{\mathrm{RF}}$ (as sums of squares of the responses, and of the fitted values under each model, all 3 being corrected by subtracting $n\overline{Y}^2$), then the other sums of squares by subtraction.

That is, $SS^\star_{\mathrm{T}} = \boldsymbol{Y}^T\boldsymbol{Y} - n\overline{Y}^2$ and

$$SS^\star_{\mathrm{RR}} = \widehat{\boldsymbol{Y}}_1^T\widehat{\boldsymbol{Y}}_1 - n\overline{Y}^2 \quad SS^\star_{\mathrm{RF}} = \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}} - n\overline{Y}^2$$

where $\widehat{\boldsymbol{Y}}_1$ and $\widehat{\boldsymbol{Y}}$are the fitted values from the reduced and full models.

The anova() function in **R** does all the work for you. Note that it displays the regression sums of squares for the model with just the 1st term, then the extra attributable to each additional term, and finally the residual sum of squares. That is, numbering the models in order of the terms included as $1, 2, \ldots, m$, the displayed sums of squares are
$SS^\star_{\mathrm{R}1}, SS^\star_{\mathrm{R}2} - SS^\star_{\mathrm{R}1}, \ldots, SS^\star_{\mathrm{R}m} - SS^\star_{\mathrm{R}(m-1)}, SS_{\mathrm{E}m}$
The $F$ tests performed by the anova() function thus relate to the successive inclusion of each term sequentially.
The order in which terms are included is therefore important. The only exception to this is when the terms in the linear model are *orthogonal*.

# Orthogonality

Suppose that $X$ can be partitioned as $X = [X_1|X_2|\cdots|X_t]$ where $X_i^T X_j = 0$ for all $i \neq j$, each $X_i$ representing a block of $p_i \geq 1$ columns. Then

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 & \cdots & X_1^T X_t \\ X_2^T X_1 & X_2^T X_2 & \cdots & X_2^T X_t \\ \vdots & \vdots & & \vdots \\ X_t^T X_1 & X_t^T X_2 & \cdots & X_t^T X_t \end{pmatrix}$$

So

$$(X^T X)^{-1} = \begin{pmatrix} (X_1^T X_1)^{-1} & 0 & \cdots & 0 \\ 0 & (X_2^T X_2)^{-1} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (X_t^T X_t)^{-1} \end{pmatrix}$$

and finally

$$\widehat{\beta} = (X^T X)^{-1} X^T \boldsymbol{Y} = \begin{pmatrix} (X_1^T X_1)^{-1} X_1^T \boldsymbol{Y} \\ \vdots \\ (X_t^T X_t)^{-1} X_t^T \boldsymbol{Y} \end{pmatrix}$$

Thus each block of parameters is estimated just as if the other blocks were not present in the model.
Furthermore, since $\text{var}(\widehat{\beta}) = \sigma^2(X^T X)^{-1}$, estimates in different blocks are uncorrelated.
Finally, the regression sums of squares relate simply. For a model including blocks $k \in T \subset \{1, 2, \ldots, t\}$,

$$SS_{\text{R}} = \left(\sum_{k \in T} X_k \widehat{\beta}_k\right)^T \sum_{k \in T} X_k \widehat{\beta}_k = \sum_{k \in T} \widehat{\beta}_k^T X_k^T X_k \widehat{\beta}_k$$

Taking the Full model to be any model including block $j$, and the Reduced model to be the same but excluding block $j$, then $SS_{\text{RF}} - SS_{\text{RR}} = \widehat{\beta}_j^T X_j^T X_j \widehat{\beta}_j$.

The ANOVA table can then be unambiguously constructed as

| Source | SS | df |
|---|---|---|
| Regression on $\boldsymbol{\beta}_1$ | $\widehat{\boldsymbol{\beta}}_1^T X_1^T X_1 \widehat{\boldsymbol{\beta}}_1$ | $p_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| Regression on $\boldsymbol{\beta}_t$ | $\widehat{\boldsymbol{\beta}}_t^T X_t^T X_t \widehat{\boldsymbol{\beta}}_t$ | $p_t$ |
| Residual | $SS_{\mathrm{EF}}$ | $n - p$ |
| Total, uncorrected | $SS_{\mathrm{T}}$ | $n$ |

# Sketch proof of $\chi^2$ and $F$ results

The results quoted in slides 35, 42 and 46, and many other results used for testing hypotheses in normal linear models, follow from the following proposition.

Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 I)$, and that $A_1, A_2, \ldots, A_m$ are real $n \times n$ matrices such that

- $\sum_{r=1}^m A_r = I_n$
- $A_r^T A_s = 0$ for all $r \neq s$

These are the same as $\sum_{r=1}^m A_r \boldsymbol{y} = \boldsymbol{y}$ and $(A_r \boldsymbol{y})^T (A_s \boldsymbol{y}) = 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$. It follows that $A_r^2 = A_r = A_r^T$ for all $r$. Then

- $A_r \boldsymbol{Y}$, $r = 1, 2, \ldots, m$ are independent
- $A_r \boldsymbol{Y} \sim N_n(A_r \boldsymbol{\mu}, \sigma^2 A_r)$
- $E(\boldsymbol{Y}^T A_r \boldsymbol{Y}) = \sigma^2 p_r + \boldsymbol{\mu}^T A_r \boldsymbol{\mu}$, where $p_r = \mathrm{tr}(A_r) = \mathrm{rank}(A_r)$.
- $\boldsymbol{Y}^T A_r \boldsymbol{Y} \sim \sigma^2 \chi^2_{p_r} \Leftrightarrow A_r \boldsymbol{\mu} = 0$

This can be proved using 1st year Linear Algebra, but we will omit the proof.

# Corollaries

Distribution of $\widehat{\beta}$ and $S(\widehat{\beta})$. Let $m = 2$, $A_1 = H$, $A_2 = (I - H)$; we find $p_1 = p$ and $p_2 = n - p$. Also note that $(X^T X)^{-1} X^T A_1 \boldsymbol{Y}$ simplifies to $\widehat{\beta}$, so $\widehat{\beta}$ is indeed a function of $A_1 \boldsymbol{Y}$. Details left as exercise. This also proves the results on page 42.

Full vs. Reduced model $F$ test. Let $m = 3$, $A_1 = X_1(X_1^T X_1)^{-1} X_1^T$, $A_2 = X(X^T X)^{-1} X^T - A_1$, $A_3 = I - A_1 - A_2$; we find $p_1 = q$, $p_2 = p - q$ and $p_3 = n - p$. We can simplify: $\boldsymbol{Y}^T A_2 \boldsymbol{Y} = SS_{\mathrm{ER}} - SS_{\mathrm{EF}}$ and $\boldsymbol{Y}^T A_3 \boldsymbol{Y} = SS_{\mathrm{EF}}$. Details left as exercise.

Correcting for the mean. This just corresponds to taking one of the terms, say $\boldsymbol{Y}^T A_1 \boldsymbol{Y}$, to be $n\overline{Y}^2$. The corresponding $p_1 = 1$.

# 9. Factorial experiments

This section is concerned with the situation where all of the explanatory variables are factors. The simplest examples are (e) and (f) of Section 1 (one- and two-way analysis of variance). Such data very often arises from designed experiments, rather than observational studies, and the design of these experiments is the subject of the Experimental design unit.

# Least squares estimates

As we noted before, the $X$matrix in these problems may not be full-rank, so we usually find least squares estimates from first principles, not the matrix formula.

One-way analysis. In double-subscript notation:

$$Y_{ij} = \alpha_i + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$$

where $n = \sum_{i=1}^{k} n_k$. Or we may prefer to express the effects of the factor as departures from an overall mean:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$$

Note that the least squares estimates cannot be unique: we need to impose a constraint to ensure uniqueness – we will use $\sum_{i=1}^{k} n_i\alpha_i = 0$, but other possibilities are $\alpha_1 = 0$, $\sum_i \alpha_i = 0$, etc.

The sum of squares is

$$S(\mu, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \big( (Y_{ij} - \overline{Y}_{i.}) + (\overline{Y}_{i.} - \overline{Y}_{..} - \alpha_i) + (\overline{Y}_{..} - \mu) \big)^2$$

where dots as subscripts to $\overline{Y}$ indicate averaging over missing subscripts. Expanding out the square, we find that because $\sum_{i=1}^{k} n_i \alpha_i = 0$, the cross-terms cancel, and we get:

$$S(\mu, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y}_{i.} - \overline{Y}_{..} - \alpha_i)^2$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y}_{..} - \mu)^2$$

Hence the LSE's are $\widehat{\mu} = \overline{Y}_{..}$ and $\widehat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{..}$

We also see immediately that the residual sum of squares is $SS_{\mathrm{E}} = S(\widehat{\mu}, \widehat{\alpha}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$ and the same argument gives the sum-of-squares decomposition

$$SS_{\mathrm{T}} = \boldsymbol{Y}^T \boldsymbol{Y} = SS_{\mathrm{E}} + SS_{\mathrm{group}} + n\overline{Y}_{..}^2$$

where $SS_{\mathrm{group}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{Y}_{i.} - \overline{Y}_{..})^2$
The ANOVA table is, in part:

| Source | SS | df |
|---|---|---|
| Due to groups | $SS_{\mathrm{group}}$ | $k - 1$ |
| Residual | $SS_{\mathrm{E}}$ | $n - k$ |
| Total, corrected | $SS_{\mathrm{T}}^{\star}$ | $n - 1$ |

and the one-way analysis of variance test of the hypothesis that all the $\alpha_i$ are equal to 0 refers

$$F = \frac{SS_{\mathrm{group}}/(k-1)}{SS_{\mathrm{E}}/(n-k)}$$

to the $F_{k-1,n-k}$ distribution (justified, again, by the result on slide 51).

Row-and-column analysis. We can follow the same line with the row-plus-column model (example (f))

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, i = 1, 2, \ldots, r; j = 1, 2, \ldots, c$$

but instead we will look at the more general case:
Two-way analysis, with replication. Suppose that we have an *equal* number $\ell$ of repeat or replicate observations in each $(i, j)$ cell of the two-way layout. The model is as above, but with an additional term:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, \ldots, r; j = 1, 2, \ldots, c; k = 1, 2, \ldots, \ell$. Let $n = rc\ell$. As in the previous analysis, least squares estimates are not unique as the model stands, and we need to impose some constraints. We assume $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, and $\sum_i \gamma_{ij} = 0$ for all $j$ and $\sum_j \gamma_{ij} = 0$ for all $i$.

The sum of squares is

$$S(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{\ell} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{\ell} \Big((Y_{ijk} - \overline{Y}_{ij.}) + (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...} - \gamma_{ij})$$

$$+ (\overline{Y}_{i..} - \overline{Y}_{...} - \alpha_i) + (\overline{Y}_{.j.} - \overline{Y}_{...} - \beta_j) + (\overline{Y}_{...} - \mu)\Big)^2$$

As in the one-way analysis, the choice of constraints means that all cross-terms cancel when the square is multiplied out, so that $S(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the triple sum (over $i$, $j$ and $k$) of $(Y_{ijk} - \overline{Y}_{ij.})^2$, $(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...} - \gamma_{ij})^2$, $(\overline{Y}_{i..} - \overline{Y}_{...} - \alpha_i)^2$, $(\overline{Y}_{.j.} - \overline{Y}_{...} - \beta_j)^2$, and $(\overline{Y}_{...} - \mu)^2$.
The least squares estimates are therefore

$$\widehat{\gamma}_{ij} = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...} \quad , \quad \widehat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{...}$$

$$\widehat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{...} \quad \text{and} \quad \widehat{\mu} = \overline{Y}_{...}$$

The residual sum of squares is
$SS_{\mathrm{E}} = S(\widehat{\mu}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{\ell} (Y_{ijk} - \overline{Y}_{ij.})^2$ and the sum-of-squares decomposition is

$$SS_{\mathrm{T}} = \boldsymbol{Y}^T \boldsymbol{Y} = SS_{\mathrm{E}} + SS_{\mathrm{row}} + SS_{\mathrm{col}} + SS_{\mathrm{inter}} + n\overline{Y}^2_{...}$$

where $SS_{\mathrm{row}} = c\ell \sum_{i=1}^{r} (\overline{Y}_{i..} - \overline{Y}_{...})^2$, $SS_{\mathrm{col}} = r\ell \sum_{j=1}^{c} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$, and $SS_{\mathrm{inter}} = \ell \sum_{i=1}^{r} \sum_{j=1}^{c} (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$.

ANOVA table:

| Source | SS | df |
|---|---|---|
| Rows | $SS_{\text{row}}$ | $r-1$ |
| Columns | $SS_{\text{col}}$ | $c-1$ |
| Interaction | $SS_{\text{inter}}$ | $(r-1)(c-1)$ |
| Cells | $SS_{\text{cells}}$ | $rc-1$ |
| Residual | $SS_{\text{E}}$ | $rc(\ell-1)$ |
| Total, corrected | $SS_{\text{T}}^{\star}$ | $n-1$ |

where $SS_{\text{cells}} = SS_{\text{row}} + SS_{\text{col}} + SS_{\text{inter}}$ of course. By analogy with previous cases, this gives $F$ tests of hypotheses that there is no interaction ($\gamma_{ij} \equiv 0$), or no row effects ($\alpha_i \equiv 0$) or no column effects ($\beta_j \equiv 0$). As usual, the $F$ ratios and their degrees of freedom are read off from the table; for example the test for no interaction is based on referring

$$F = \frac{SS_{\text{inter}}/((r-1)(c-1))}{SS_{\text{E}}/(rc(\ell-1))} \quad \text{to} \quad F_{(r-1)(c-1),rc(\ell-1)}$$

# Verification of assumptions of ANOVA theorem

Here we spell out the checking of assumptions of the ANOVA theorem for one case – the two-way replicated analysis.

The response vector $\boldsymbol{Y}$ and its sum of squares are decomposed into $m = 5$ pieces. The matrices $A_1, A_2, \ldots, A_5$ are defined implicitly by their effect on the vector $\boldsymbol{Y}$. For $r = 1, 2, \ldots, 5$, the vector $A_r\boldsymbol{Y}$ is indexed by a triple subscript $(i, j, k)$, like the response vector itself. The $(i, j, k)$ entries of $A_r\boldsymbol{Y}$, for $r = 1, 2, \ldots, 5$ are, respectively: $(Y_{ijk} - \overline{Y}_{ij.})$, $(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})$, $(\overline{Y}_{i..} - \overline{Y}_{...})$, $(\overline{Y}_{.j.} - \overline{Y}_{...})$, and $(\overline{Y}_{...})$.

The assumption that $\sum_r A_r = I$ is verified by observing that these add up to $Y_{ijk}$. The assumption that $A_r^T A_s = \mathbf{0}$ is verified by taking any two of these 5 terms, summing over all $(i, j, k)$ and checking that the result is 0 (using the given constraints).

For example

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{\ell}(\overline{Y}_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})(\overline{Y}_{i..}-\overline{Y}_{...})$$

$$=\ell\sum_{i=1}^{r}(\overline{Y}_{i..}-\overline{Y}_{...})\sum_{j=1}^{c}(\overline{Y}_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})$$

$$=\ell\sum_{i=1}^{r}(\overline{Y}_{i..}-\overline{Y}_{...})(c\overline{Y}_{i..}-c\overline{Y}_{i..}-c\overline{Y}_{...}+c\overline{Y}_{...})$$

$$=\ell\sum_{i=1}^{r}(\overline{Y}_{i..}-\overline{Y}_{...})0=0$$

# Motivation for the model with interaction

A feature of the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$; $k = 1, 2, \ldots, \ell$, or in **R**: `Y ~ Row*Column` is that it allows us to examine whether row and column effects are *additive* or not. For example, consider agricultural trial data where rows represent sites and columns reprsent varieties of a crop. The additive model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

(`Yield~Site+Variety`) assumes that the advantage of a good variety affects the yield of crop at all sites equally. It is perhaps more realistic to accept the possibility that this is not true, that is to allow the $\gamma_{ij}$term into the model, which becomes `Yield~Site*Variety`. The additive model, without the interaction term, is however easier to interpret, so testing for the presence of the interaction (with the $F$ test on slide 58) would usually be the first inference to be performed. If the hypothesis of no interaction is not rejected, then we can use the simpler interpretation.

Whether or not interaction is detected depends on the scale on which the measurements are made. Sometimes we consider transforming the responses using some monotonic function, to achieve additivity. For example, imagine the data were generated, without noise, from the model

$$Y_{ijk} = \alpha_i \times \beta_j$$

(a multiplication table!) Then an additive model would not fit very well. If we had replicated data with a little noise, and we tested for interaction, we would probably conclude that it was indeed present. However if we take logs, we get

$$\log Y_{ijk} = \log \alpha_i + \log \beta_j$$

and the additive model would fit perfectly.
Note however that taking transformations like this alters the distributional shape: it is not possible for both $Y$and $\log Y$ to be normally distributed.

# Non-replicated case

When there is only $\ell = 1$ observation in each $(i, j)$ cell, we do not need the subscript $k$and the model becomes

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, i = 1, 2, \ldots, r; j = 1, 2, \ldots, c.$$

We can immediately see a problem: we will not be able to distinguish the interaction $\gamma_{ij}$from the noise $\epsilon_{ij}$. This is an inevitable consequence of having only one observation per cell.
If you ignored this, then the residual sum of squares $SS_{\mathrm{E}}$on slide 57, namely $\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{\ell} (Y_{ijk} - \overline{Y}_{ij.})^2$ collapses to zero, and its degrees of freedom $rc(\ell - 1)$ are also zero. If you then tried to perform the $F$tests, the denominator mean square would be $0/0$, not defined.

We have to assume there is nointeraction, and omit that term, giving

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, i = 1, 2, \ldots, r; j = 1, 2, \ldots, c.$$

which is just example (f) again. Correspondingly we do not break down the sum of squares into so many pieces. With $SS_{\text{row}}$ and $SS_{\text{column}}$ defined as on slide 57, but with the thirdsubscript (always ".") omitted, we put $SS_{\text{E}} = \sum_{i=1}^{r} \sum_{j=1}^{c} (\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$, and obtain

$$SS_{\text{T}} = \boldsymbol{Y}^T \boldsymbol{Y} = SS_{\text{E}} + SS_{\text{row}} + SS_{\text{col}} + n\overline{Y}_{..}^2$$

The ANOVA table simplifies to

| Source | SS | df |
|---|---|---|
| Rows | $SS_{\text{row}}$ | $r - 1$ |
| Columns | $SS_{\text{col}}$ | $c - 1$ |
| Residual | $SS_{\text{E}}$ | $(r - 1)(c - 1)$ |
| Total, corrected | $SS_{\text{T}}^{\star}$ | $n - 1$ |

It is important to remember that we have made an assumption (that row and column effects are additive), and make some effort to assess if that is supported by the data, even if the $F$ test is not possible. The diagnostic plots from Section 3 should reveal any problem – if there is interaction, it will usually show up in the fitted values/residuals plot.

# Unequally-replicated case

Sometimes, by design or accident (e.g. missing data), our data has an unequal degree of replication:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, \ldots, r; j = 1, 2, \ldots, c; k = 1, 2, \ldots, \ell_{ij}$, where the $\ell_{ij}$ are not all equal. The total number of observations is $n = \sum_i \sum_j \ell_{ij}$.
With appropriate changes (the $\ell_{ij}$ go inside the summations in the definitions of the sums of squares, and the residual degrees of freedom becomes $\sum_i \sum_j (\ell_{ij} - 1)$), the ideas in slides 56 to 58 mostly still apply, except that we lose orthogonality, e.g.

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \ell_{ij} (\overline{Y}_{i..} - \overline{Y}_{...})(\overline{Y}_{.j.} - \overline{Y}_{...})$$

is not in general zero. The ANOVA then becomes more complicated, and must be interpreted sequentially, as in the case of general regression.

# Blocks and treatments

Suppose we want to compare a set of *treatments*, for example, drug therapies, fertilisers or varieties in agriculture, industrial process settings, or teaching methods, and we want to make our comparisons on a broad range of experimental units, for example, patients of different ages, fields in different locations, etc. Groups of units that are similar are called *blocks*, and the standard arrangement is the randomised block design, where we compare say $r$different treatments, applying each to $c$different blocks of $r$ units, the units being randomly assigned to the treatments in each block in order to protect against any accidental bias. Sometimes there are replicates, e.g. there might be $\ell r$ units in each block, with $\ell$ assigned to each treatment.
In this situation, a two-way analysis of variance is appropriate, even though we are typically only interested in differences between treatments, not between blocks.

But you might think that a one-way analysis of variance would be sufficient. Does it matter which you do?

Suppose we have unreplicated data and that the response to treatment $i$ in block $j$ is $Y_{ij}$.

The $F$ test of the hypothesis of no treatment effect refers

$$F = \frac{SS_{\mathrm{row}}/(r-1)}{SS_{\mathrm{E}}/((r-1)(c-1))} \quad \text{to} \quad F_{(r-1),(r-1)(c-1)}$$

according to slide 64. If we did a one-way analysis test of the same hypothesis we would refer

$$F^{[1]} = \frac{SS_{\mathrm{row}}/(r-1)}{SS_{\mathrm{E}}^{[1]}/r(c-1)} \quad \text{to} \quad F_{(r-1),r(c-1)}$$

from slide 55, changing the notation there to match the present situation.

The numerators in the $F$ ratios are the same, but the denominators are different. The residual sums of squares are $SS_{\mathrm{E}} = \sum_{i=1}^{r} \sum_{j=1}^{c} (\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$ and $SS_{\mathrm{E}}^{[1]} = \sum_{i=1}^{r} \sum_{j=1}^{c} (Y_{ij} - \overline{Y}_{i.})^2$ from slides 64 and 55 respectively, and the degrees of freedom are correspondingly different. If there are substantial differences between block means, then $SS_{\mathrm{E}}^{[1]} \gg SS_{\mathrm{E}}$. The one-way test is then much less sensitive (less likely to reject the null hypothesis, even if it is false). The distinction is the same as that between the paired comparison $t$-test and the two-sample $t$-test, which is what this all reduces to if $r = 2$.

# Connection with *t* tests

A point about connections between $t$ and $F$ tests was made in the regression context in question 1 of sheet 4, but it is more general.

The percentage points in the tables satisfy $t_\nu(\alpha/2)^2 = F_{1,\nu}(\alpha)$ for all $\nu, \alpha$, or in **R**, `qt(1-p/2,nu)^2=qf(1-p,1,nu)` for all `(p,nu)`, and recalling the definitions of the $t$ and $F$ distributions in terms of independent normal and $\chi^2$ random variables, it is not hard to see why. As for the test statistics, consider the non-replicated two-way analysis, with $r = 2$. Let $d_j = Y_{1j} - Y_{2j}$. Then $SS_{\mathrm{row}} = c\sum_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$ can be written as $(c/2)\overline{d}^2$, while $SS_{\mathrm{E}} = \sum_i\sum_j(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} - \overline{Y}_{..})^2$ is the same as $(1/2)\sum_j(d_j - \overline{d})^2$. So

$$F = \frac{SS_{\mathrm{row}}/(r-1)}{SS_{\mathrm{E}}/((r-1)(c-1))} = \frac{c\overline{d}^2}{\sum_j(d_j - \overline{d})^2/(c-1)}$$

which is clearly just the square of the paired-comparison $t$ statistic.

# 10. Residual analysis

Residual analysis allows us to

1. check the assumptions of the model,
2. check the adequacy of the model,
3. detect outliers.

The residuals are given by

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\beta}.$$

Before examining the residuals, it is common to standardise them as follows

$$r_i = \frac{\hat{e}_i}{\widehat{\sigma}_{e_i}}$$

where $\widehat{\sigma}_{e_i}$ is the estimated standard error for $\hat{e}_i$. Then under the assumptions of the NLM, the $r_i$ should be **i.i.d. N(0, 1)** random variables. We can check this using graphical methods.

# Normality assumption

The normality assumption can be checked using a q-q plot. Plot $r_{(i)}$, the $i$th rank order residuals against

$$\Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$$

the “expected normal deviate”. If the $r_i$ are normally distributed, the points should fall on a straight line at a 45 degree angle.

# Constant variance assumption

Plot $r_i$ versus fitted values $\hat{y}_i$ and look for systematic changes in spread about $r = 0$.

# Model adequacy

We need check if there are any systematic trends in the residuals.
Plot $r_i$ versus explanatory variables in the model.
Plot $r_i$ versus explanatory variables **not** included in the model.
Plot $r_i$ versus $\widehat{y}_i$.
Plot $r_i$ versus time (if applicable).

# Outliers

If the residuals have been standardised we expect the majority to fall in the range (-2, 2). We can check the plot of $r_i$ vs. $y_i$ for any observations with "large" $r_i$. Such observations should be checked.

# Variable Selection

Generally aim for a **parsimonious** model: the simplest model that describes the data well. Need to balance competing objectives:
*Include as many variables as possible*

- avoid bias in $\widehat{\beta}_j$'s
- include all variables with predictive power

*Include as few variables as possible*

- reduce variance of $\widehat{y}_i$'s
- keep model simple

# Leverage

Recall

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta},$$

So

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

where $H$ is called the **hat matrix**. It can be shown that residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ have $E(\hat{\boldsymbol{e}}) = 0$ and $Var(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, or

$$Var(\hat{\epsilon}_i) = \sigma^2(1 - H_{ii}) = \sigma^2(1 - h_i)$$

where $h_i$ is known as the **leverage**. An observation with high leverage, $h_i$, makes $Var(\hat{\epsilon}_i)$ small, forcing the fit close to $y_i$.
The average of all $n$ leverages is $k/n$, so anything with leverage $> 2k/n$ should be considered more closely.

# Cook's Distance

**Cook's Distance** is a measure of **influence**. It can be defined as

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\beta}_{(i)} - \widehat{\beta})}{k\widehat{\sigma}^2}$$

where $\widehat{\beta}_{(i)}$ is the estimate of $\widehat{\beta}$ omitting observation $i$.
It can be shown that

$$D_i = \frac{\hat{\epsilon}^2}{k\sigma^2}\left(\frac{h_i}{(1-h_i)^2}\right)$$

showing that $D_i$ is a function of the residual and the leverage.
Observations with $D_i > 0.5$ may be of concern and observations with $D_i > 1$ should definitely be checked.

# Multicollinearity

Since

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

in general all $\widehat{\beta}_j$ are correlated and each $\widehat{\beta}_j$ changes on addition/deletion of variables.

When there are near linear dependencies between variables, the variables are described as **mulitcollinear**. So

- variances of parameter estimates are high
- parameter estimates change drastically on addition/deletion of variables
- danger of hidden extrapolation

# Diagnosing Multicollinearity

Can look at pairwise scatterplots, but this only indicates pairwise correlation.
Multicollinearity can be detected by **variance inflation factors** $VIF$.
It can be shown that

$$Var(\widehat{\beta}_j) = \frac{1}{1 - R_j^2}\left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_{ji} - \overline{x}_j)^2}\right)$$

or more simply

$$VIF_j = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the $R^2$ value for the regression of $x_j$ on all the other regression variables.
A $VIF > 10$ suggest serious multicollinearity and one should be concerned if $VIF > 5$.

# 11. Variable Selection

Variables can often be selected manually using a sensible model building strategy, e.g.

- starting from a full model then simplifying
- including variables on which response is assumed to depend then considering the addition of further variables

However, when there are many variables this becomes more difficult. One approach is to use automatic selection procedures to suggest a model/candidate models.

# Forward Selection

Start with intercept then add variables one at a time.
Suppose the current model has residual SS

$$SSE_{(1)}$$

We try adding each of the remaining variables and find the variable $x_j$ which when added gives the model with smallest residual SS

$$SSE_{(2)}$$

Under $H_0 : \beta_j = 0$,

$$F_0 = \frac{SSE_{(1)} - SSE_{(2)}}{\widehat{\sigma}^2} \sim F_{1,n-k}$$

where $\widehat{\sigma}^2$ is the residual mean square of the full model with $k$ parameters. If $F_0 > F_{1,n-k;\alpha}$, $x_j$ is added and the next variable is sought, otherwise we stop.

# Backward Elimination

Start with full model then drop variables one at a time.
Suppose the current model has residual SS

$$SSE_{(1)}$$

We try dropping each of the remaining variables and find the variable $x_j$ which when dropped gives the model with smallest residual SS

$$SSE_{(2)}$$

Under $H_0; \beta_j = 0$

$$F_0 = \frac{SSE_{(2)} - SSE_{(1)}}{\widehat{\sigma}^2} \sim F_{1,n-k}$$

If $F_0 < F_{1,n-k;\alpha}$, $x_j$ is dropped from the model and the next variable to be dropped is sought, otherwise stop.

# Stepwise Selection

With forward selection, a variable include early on may become unnecessary after a latter variable is added.

With backward elimination, a variable removed early on may become significant after others are deleted.

Stepwise selection is a combination of the two: proceed as with forward selection, but at each stage test to see if any variables currently in the model can be dropped.

Stop when all potential deletions would significantly increase the residual sum of squares and when all potential additions would not significantly decrease the residual sum of squares.

# Problems with Sequential Procedures

- do not guarantee the “best” model will be chosen
- all select one model - may be several good models
- different procedures may select different models
- order in which variables enter/leave does not reflect “importance”
- multicollinearity is not diagnosed
- multiple significance testing - $\alpha$ not true level of significance

# All subsets regression

If the number of variables is not too large (say $< 30$) it is possible to consider all possible regressions and choose “the best” according to some criteria. R$^2$ and R$^2_{adj}$ are two possible criteria. Another possibility is Mallow’s $C_p$ statistic

$$C_p = \frac{SSE_p}{\widehat{\sigma}^2} + 2p - n$$

where $p$ is the number of parameters in the current model, $SSE_p$ is the corresponding residual SS and $\widehat{\sigma}^2$ is the residual mean square from the full model.

If a model with $p$ parameters is adequate there will be no lack of fit or bias, so we can estimate residual variance as

$$\frac{SSE_p}{(n-p)} \Rightarrow \qquad E(SSE_p) = (n-p)\sigma^2 \Rightarrow E(C_p) = p$$

Therefore select model with fewest parameters that has $C_p \approx p$. Usually, but not always the lowest $C_p$.

All subsets regression is carried out in R using `leaps()` from the `leaps` package.

# AIC Model Selection Procedures

The Akaike information criterion, $AIC$, is given by

$$AIC = -2l(\hat{\beta}) + 2k$$

where

- $l(\hat{\beta})$ the log-likelihood of the candidate model given the data when evaluated at the maximum likelihood estimate (MLE) $\hat{\beta}$;
- $k$ is the number of estimated parameters in the candidate model

The $AIC$ in isolation is meaningless. Rather, this value is calculated for every candidate model and the *best* model is the candidate model with the *smallest* $AIC$, say $AIC^*$.

# BIC Model Selection Procedures

The Bayesian information criterion *BIC*, also referred to as the Schwarz information criterion is another model selection criterion is

$$BIC = -2l(\hat{\beta}) + k \log n$$

It is based on information theory but set within a Bayesian context. The difference between the *BIC* and the *AIC* is the greater penalty imposed for the number of parameters by the former than the latter.

# BIC Model Selection Procedures

The best model is the one that provides the minimum $BIC$, say $BIC^{*}$
Given $M$ models, $\Delta BIC = BIC - BIC^{*}$ can be interpreted as evidence *against a candidate model being the best model*. The rules of thumb for $\Delta BIC$ are:

- $< 2$, it is not worth more than a bare mention.
- $[2, 6]$, the evidence against the candidate model is positive.
- $[6, 10]$, the evidence against the candidate model is strong.
- $> 10$, the evidence is very strong.

Similarly for $\Delta AIC = AIC - AIC^{*}$.

# Problems with Automatic Selection Procedures

All subsets regression, like the sequential procedures has the attraction of selecting the “best” model.
However none of these procedues takes into account other factors, e.g.

- subject knowledge
- costs
- examination of residuals
- multicollinearity

All methods of variable selection can overstate significance of variables in the final model.

# 12. Weighted Least Squares

The method of weighted least squares (WLS) can be used when the ordinary least squares assumption of constant variance in the errors is violated (heteroscedasticity). The model is

$$E[Y|\boldsymbol{X}] = \boldsymbol{X}\beta$$

with

$$Cov(Y|\boldsymbol{X}) = \sigma^2 \boldsymbol{W}^{-1}$$

or

$$var(Y|\boldsymbol{X} = x_i) = var(e_i) = \sigma^2/w_i, \; w_i > 0$$

where $\boldsymbol{W}$ is a diagonal positive definite matrix.
Show that $\hat{\beta}_w = (\boldsymbol{X}' W \boldsymbol{X})^{-1} \boldsymbol{X}' W \boldsymbol{y}$ is unbiased and that

$$var(\hat{\beta}_w) = \sigma^2 (\boldsymbol{X}' W \boldsymbol{X})^{-1}$$

# Weighted Least Squares: Exercise

Show that the WLS estimator that minimises

$$SSE_w(\beta) = (Y - \boldsymbol{X}\beta)' W (Y - \boldsymbol{X}\beta)$$

is

$$\hat{\beta}_w = (\boldsymbol{X}' W \boldsymbol{X})^{-1} \boldsymbol{X}' W \boldsymbol{y}.$$

and prove that $\hat{\beta}_w$ is unbiased and that

$$var(\hat{\beta}_w) = \sigma^2 (\boldsymbol{X}' W \boldsymbol{X})^{-1}$$

with $\sigma^2 = SSE_w/d$, $d = n - k$ d.f.

# Dependence and heteroscedasticity

### Exercise Dependence and heteroscedasticity

*Suppose that* $\boldsymbol{Y}|\boldsymbol{X}$ *is* $MVN(\boldsymbol{X}\beta, \sigma^2 W)$ *where* $W$ *is a known positive definite matrix. Show that the MLE of* $\boldsymbol{\beta}$ *is given by the generalised least squares estimator*

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}' W^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' W^{-1} \boldsymbol{y}.$$

# How to choose the weights

Given that

$$\hat{\beta}_w = (\boldsymbol{X}' W \boldsymbol{X})^{-1} \boldsymbol{X}' W \boldsymbol{y} \text{ and } var(Y|\boldsymbol{X} = x_i) = var(e_i) = \sigma^2 / w_i, \; w_i > 0$$

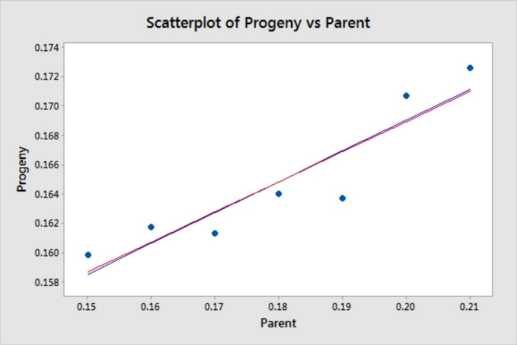
We can define each $w_i = 1/\sigma_i^2$ and since

- each weight is inversely proportional to the error variance, it reflects the information in that observation;
- an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance;
- weights have to be known (or are usually estimated) up to a proportionality constant.

# How to choose the weights: example

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Parent | 0.21 | 0.2 | 0.19 | 0.18 | 0.17 | 0.16 | 0.15 |
| Progeny | 0.1726 | 0.1707 | 0.1637 | 0.164 | 0.1613 | 0.1617 | 0.1598 |
| $\sigma_i$ | 0.01988 | 0.01938 | 0.01896 | 0.02037 | 0.01654 | 0.01594 | 0.01763 |

`Progeny = 0.127 +0.21Parent` for simple linear regression with $w_i = 1$.
`Progeny = 0.128 +0.20Parent` for $w_i = 1/\sigma_i^2$ and we get the weighted LS.



The *red* line is the WLS, the *blue* line is the LS.

# How to choose the weights

Other cases where the weights might be known:

- If the i-th response is an average of $n_i$ equally variable observations, then $w_i = n_i$;
- If the i-th response is a total of $n_i$ observations, then $w_i = 1/n_i$;
- If the variance is proportional to some predictor $x_i$, then $w_i = 1/x_i$.

# How to choose the weights

- If a residual plot against a predictor exhibits a megaphone shape, then regress the absolute values of the residuals against that predictor. The resulting fitted values $\hat{y}^*$ of this regression are estimates of $\sigma_i$, so $w_i = 1/\sigma_i^2$;
- If a residual plot against the fitted values exhibits a megaphone shape, then regress the absolute values of the residuals against the fitted values. So the resulting $\hat{y}^*$ are estimates of $\sigma_i$.
- If a residual plot of the squared residuals against a predictor exhibits an upward trend, then regress the squared residuals against that predictor. So the resulting $\hat{y}^*$ are estimates of $\sigma_i$.
- If a residual plot of the squared residuals against the fitted values exhibits an upward trend, then regress the squared residuals against the fitted values. So the resulting $\hat{y}^*$ are estimates of $\sigma_i$.

After using one of these methods to estimate $w_i = 1/\sigma_i^2$, then use these weights in estimating a WLS regression model.

# Box-Cox Transformations

Try stabilising the variance using **Box-Cox transformation**:

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda\text{ne}0 \\ \log y & \lambda = 0 \end{cases}$$

Assuming all regression parameters are set at their maximum likelihood estimates the **profile likelihood** for $\lambda$ is

$$L(\hat{\lambda}) = \text{const} - \frac{n}{2} \log \text{SSE}(z^{(\lambda)})$$

where $z^{(\lambda)} = y^{(\lambda)}/\dot{y}^{(\lambda-1)}$, where $\dot{y}$ represents the geometric mean, and SSE is the residual sum of squares for the regression of $z^{(\lambda)}$.

# Problems shown by residuals

Heavy-tailed distribution, variance increases with $\widehat{y}$
$\rightarrow$ transformation (Box-Cox)
Positively skewed distribution
$\rightarrow$ counts? log-linear modelling
Variance shows “double bow” shape
$\rightarrow$ proportions? logistic regression
Relationship with $x_j$
$\rightarrow$ include $x_j$ in the model
Increasing variance with $x_j$
$\rightarrow$ weighted least squares
Outliers
$\rightarrow$ check validity; check influence

# 13. Data collection in finite samples

How can we collect data so that the sample represents the population?

What do we even means by “represents”?

e.g. the proportion of the sample whose value lies in some interval is “like” the proportion of the population whose value lies in that interval

How can we collect data so that the sample represents the population?

Let us look at some methods of collection:

- Collecting Data with a Sample Survey
- Collecting Data with an Experiment
- Collecting Data with an Observational Study

# 13. Data collection in finite samples

How can we collect data so that the sample represents the population?

What do we even means by “represents”?

e.g. the proportion of the sample whose value lies in some interval is “like” the proportion of the population whose value lies in that interval

How can we collect data so that the sample represents the population?

Let us look at some methods of collection:

- Collecting Data with a Sample Survey
- Collecting Data with an Experiment
- Collecting Data with an Observational Study

# Collecting Data with a Sample Survey

**Definition – Simple random sample**

A simple random sample of $n$ subjects from a population is one in which each possible sample of size $n$ has the same probability (chance) of being selected.

In some sense this is the “perfect” way to ensure representativeness - certainly it is the simplest! But it is not always possible – e.g. how could you draw a random sample of 18–25 year-old people in Cameroon? On the other hand, if you make repeated measurements of a constant physical property with the same instrument, then it is reasonable to assume that you get a simple random sample.

# Some other kinds of random sampling

**Definition – Stratified random sample**

The population is divided into separate groups (“strata”) and then select a simple random sample from each stratum. This is useful for comparing groups on some variable when a particular group is relatively small and may not be adequately represented in a simple random sample

**Definition – Clustered random sample**

The population is divided into a large number of clusters, such as city blocks, and a simple random sample of the clusters is collected. This is useful when a complete listing of the population is not available.

# Collecting Data with an Experiment

Some studies use a planned experiment to generate data. An experiment compares subjects on a response variable under different conditions. Those conditions, which are levels of an explanatory variable, are called treatments. For instance, the treatments might be different drugs for treating some illness, compared in a clinical trial.

# Planning an Experiment

The researcher specifies a plan for how to assign subjects to the treatments, called the experimental design. Good experimental designs use randomization to determine which treatment a subject receives. This reduces bias and allows us to use statistical inference.

# Collecting Data with an Observational Study

In many application areas, it is not possible to conduct experiments to answer the questions of interest. We cannot randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income or usage of guns. Many studies merely observe the outcomes for available subjects on the variables of interest, without any experimental control of the subjects. Such studies are called observational studies.